Data Appendix to "The impact of weather events on schooling outcomes"

Katya Garcia-Israel

Contents

```
## Warning: package 'summarytools' was built under R version 3.6.3
```

1 Appendix description

This data appendix details the steps of collecting, processing, and merging the data needed to create the complete data set showing information about schooling and weather events. It concludes with a discussion of patterns found in the data.

The datasets used directly by the final analysis are saved in `processed-data/` at the end of this file.

2 Raw data

This section documents the datasets used in this analysis.

2.1 Household Survey data

Citation: .
Instituto Nacional de Estadistica. (2005). Encuesta de Hogares (2005) [SAV file]. Retrieved from https://www.ine.gob.bo/index.php/banco/base-de-datos-sociales

Date Downloaded: Downloaded September 2019. **Filename(s):** raw_data/filename.csv **Unit of observation:** Unit of observation at the individual level, but each individual is identified mainly by the household id. **Dates covered:** 2005

2.1.1 To obtain a copy

This household survey can be found on the Bolivian National Statistical Institute website, under the menu for 'Banco de Datos' and 'Base de datos de encuestas sociales.' Then the household surveys may be found for the years 2005-2018, excepting 2010.

The original format of this data was from the sav file, but I imported it into Stata before importing the Stata format into R.

2.1.2 Variable descriptions

- **hh_id:** The household ID, which is the same for each member of the household. It is made up of letters and numbers.
- **nro1:** The number of the person in each household.
- **assist:** Binary variable, whether or not the individual assists school
- **urb_rur:** Binary variable, whether the household is in an urban or rural area
- **reason:** the reason why the individual is not assisting school if they are not
- **members:** number of household members excluding servants
- **sex:** the sex of the individual, Male = 1, Female =2
- **age:** How old the individual is, in years
- **relation:** The relationship of the individual to the head of household
- **literacy:** Binary variable, whether a person can read and write
- **register:** Binary variable, whether an individual registered for school or not
- **grade:** For what grade is the person registered
- **repeated:** whether the person has repeated the grade they are currently registered in
- **dept:** the department in which the person lives
- **income:** the per capita monthly income for each household, in bolivianos

2.1.3 Data import code and summary

```r
data_2005 <- read_dta(file.path("Raw data","seminar_paper_data","2005_hhsurvey.dta")) %>%
  rename(hh_id = folio,
         assist = s3_08,
         reason = s3_09,
         members = nm_y,
         sex = s1_02,
         age = s1_03,
         relation = s1_05,
         literacy = s3_01,
         register = s3_04,
         grade = s3_05b,
         repeated = s3_07,
         income = y_perc1A)

hh_2005 <- data_2005 %>%
  select(hh_id, assist, reason, members, sex, age,
         relation, literacy, register, grade, repeated, depto, urb_rur, income) %>%
```

```
  mutate(year = 2005,
         dept = case_when(depto == "1" ~ "Chuquisaca",
                          depto == "2" ~ "La Paz",
                          depto == "3" ~ "Cochabamba",
                          depto == "4" ~ "Oruro",
                          depto == "5" ~ "Potosi",
                          depto == "6" ~ "Tarija",
                          depto == "7" ~ "Santa Cruz",
                          depto == "8" ~ "Beni",
                          depto == "9" ~ "Pando"))
```

```
export_summary_table(dfSummary(hh_2005))
```

2.2 Flood data

Citation: GeoSinager. (2016). Mapa Recurrencias de Inundaciones Gestiones 2002-2012. [CSV file]. Retrieved from http://geosinager.defensacivil.gob.bo/maps/226

Date Downloaded: October 2019 **Filename(s):** raw_data/filename.csv **Unit of observation:** Municipality
Dates covered: 2002-2012

2.2.1 To obtain a copy

This data can be found at the following website: http://geosinager.defensacivil.gob.bo/maps/226. It is associated with a map showing the flood occurrences information spatially, and to get the data in a non-spatial format, the flood layer should be downloaded. It may be downloaded in a variety of different formats, including csv.

2.2.2 Variable descriptions

- **dept:** department name
- **A_2005:** how many floods occurred in 2005 in that location
- **A_2006:** how many floods occurred in 2006 in that location
- **A_2007:** how many floods occurred in 2007 in that location
- **A_2008:** how many floods occurred in 2008 in that location
- **A_2009:** how many floods occurred in 2009 in that location
- **A_2010:** how many floods occurred in 2010 in that location
- **A_2011:** how many floods occurred in 2011 in that location
- **A_2012:** how many floods occurred in 2012 in that location

2.2.3 Data import code and summary

```
flood <- read_csv(file.path("Raw data", "recurrencia_inundacion2002_2012bol.csv")) %>%
  rename(dept = NOM_DEP)
```

```
## Parsed with column specification:
## cols(
##   .default = col_double(),
```

```
##   FID = col_character(),
##   the_geom = col_character(),
##   NOM_DEP = col_character(),
##   NOM_PROV = col_character(),
##   NOM_MUN = col_character(),
##   Nivel_Rec2 = col_character()
## )

## See spec(...) for full column specifications.
```

```
flood_summary <- flood %>%
  group_by(dept) %>%
  summarize_at(vars(starts_with("A_")), sum) %>%
  pivot_longer(-dept, names_to = "year", names_prefix = "A_", values_to = "floods") %>%
  mutate(year = as.numeric(year))

export_summary_table(dfSummary(flood_summary))
```

2.3 Drought data

Citation: GeoSinager. (2016). Mapa Recurrencias de Sequias Gestiones 2002-2012. [CSV file]. Retrieved from http://geosinager.defensacivil.gob.bo/maps/225

Date Downloaded: October 2019.
Filename(s): Raw data/droughts.csv **Unit of observation:** The unit of observation is the municipality
Dates covered: 2002-2012

2.3.1 To obtain a copy

This data can be obtained from http://geosinager.defensacivil.gob.bo/maps/225. It is shown on a map with different layers, and the raw data may be downloaded in a variety of different forms, including csv.

2.3.2 Variable descriptions

- **dept:** department name
- **A_2005:** how many droughts occurred in 2005 in that location
- **A_2006:** how many droughts occurred in 2006 in that location
- **A_2007:** how many droughts occurred in 2007 in that location
- **A_2008:** how many droughts occurred in 2008 in that location
- **A_2009:** how many droughts occurred in 2009 in that location
- **A_2010:** how many droughts occurred in 2010 in that location
- **A_2011:** how many droughts occurred in 2011 in that location
- **A_2012:** how many droughts occurred in 2012 in that location

3 Data Processing and Combination

To create a complete dataset, I merged the household data with the floods data and the droughts data. Before merging the data, I cut down the datasets to include only the relevant variables needed for analysis in each file. I renamed several variables in the household survey data set to make them more understandable. I also created the year variable for the household data to ensure that it matched up correctly with the weather

data. For the flood and drought data, I grouped the occurrences by department instead of by municipality so they would match the household data, and I made the department character numeric for the merge.

```
total_2005 <- hh_2005 %>%
  left_join(flood_summary) %>%
  left_join(drought_summary)
```

```
## Joining, by = c("year", "dept")
## Joining, by = c("year", "dept")
```

```
export_summary_table(dfSummary(total_2005))

save("total_2005", file = "processed_data/analysis_data.RData")
```

4   Analysis Variables

The variables used in the final analysis are the variables from the flood and drought summaries, showing the number of occurrences in each department per year. The other variables used are from the household data, showing demographic information as well as the dependent variables related to school assistance and registration. The variables "assist" and "register" tell whether the child is currently registered and attending school, and the variables "grade" and "literacy" give further educational information showing what grade the child is in and whether they can read and write. The variables "urb_rur" and "dept" give geographic information, indicating whether the child's household is located in an urban area or a rural area and in what department of the country they are living. Demographic variables such as age, sex, relation to the head of household, number of household members, and income can be used in the analysis as controls.

5   Discussion of Data

The "assist" variable shows that around 60% of people who are registered for school currently assist, while 40% do not assist. The "reason" variable shows that two most common reasons for missing school are for work (31.3%) or vacation (20.5). Out of the people who are registered for school, 6.6% are repeating the grade they are currently in.

The patterns for households show that the average household has 5.3 members, and the average per capita monthly income per household is 258 bolivianos. Over half of the households in the sample are located in urban areas (56.3), while under half are located in rural areas (43.7). La Paz, Santa Cruz, and Cochabamba are the three departments with the most households in this sample, which corresponds to their overall population distribution.

The patterns in the weather data show that every department experienced multiple floods in 2005, while many but not all experienced droughts. The place that experienced the highest number of flood occurrences overall was Cochabamba, while Beni experienced 13 droughts in 2005.